SFU SIMON FRASER UNIVERSITY

CANSSI INCASS

Michael Smith Health Research BC

Natural Sciences and Engineering Research Council of Canada

A new variational family for Bayesian phylogenetics

Lloyd T. Elliott[1], Evan Sidrow[1], Alexandre Bouchard-Côté[2]

[1]Simon Fraser University, [2]University of British Columbia

EcoSta, Tokyo, August 2025

Genetics

SNP: Single Nucleotide Polymorphism (basepair substitution)

- COVID-19 RNA genome:
    - AUUAAAGGUUUAUACCUUCC ...
- Human DNA genome:
    - TAACCCTAACCCTAACCCTA ...

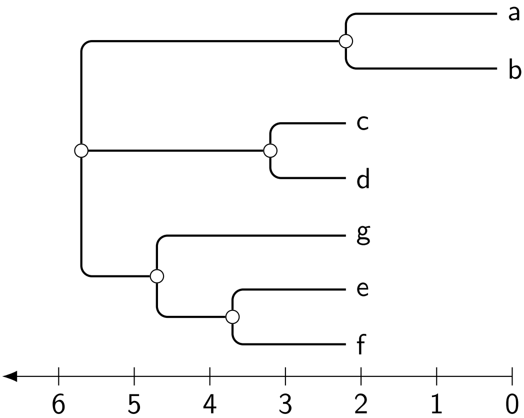
At what time did two observed genetic sequences coalesce?

- (We observe N genetic sequences)

Related problems:

- Infer mutation rates
- Discover variants of interest/variants of concern (clade emergence)
- Impute missing or ancestral sequences

Phylogenetics



- Leaf nodes: Observed RNA sequences
- Interior nodes: Unobserved
- Goal: Infer tree topology and branch lengths

Bayesian phylogenetics

- Place a prior on trees (Kingman's coalescent), develop proposals (generalized stepping-stone sampling), perform MCMC (BEAST)
  - Largest dataset studied with BEAST: ∼25k taxa (L. Lyu et al. PNAS 2025)

- Construct a variational family, perform *variational Bayes* (vB: more scalable?)
  - vB is an iterative inference algorithm that approximates the posterior using a family of functions (in contrast, MCMC approximates with a set of samples)

- Note: These methods both require a likelihood function to link the tree topology and branch lengths to the observed sequences (JC, K2P, GTR)

Single-linkage clustering

Construct a rooted tree topology from a distance matrix:

```
**Input:** Distances $\boldsymbol{T} \in \mathbb{R}_{>0}^{\binom{N}{2}}$ and taxa set $\mathcal{X} = \{\{x_1\}, \{x_2\}, \ldots, \{x_N\}\}$.
**for** $n = 1, \ldots, N-1$ **do**
    $w^*, z^* \leftarrow \arg\min_{w,z}\{t^{\{w,z\}} : w,z$ have not coalesced by the $(n-1)$-st event$\}$
    Set $W_n \in \mathcal{X}$ to be the set containing $w^*$
    Set $Z_n \in \mathcal{X}$ to be the set containing $z^*$
    $t_n \leftarrow t^{\{w^*,z^*\}}$
    Remove $W_n$, $Z_n$ from $\mathcal{X}$ and add $W_n \cup Z_n$ to $\mathcal{X}$
**end for**
$\tau \leftarrow \{\{W_n, Z_n\}\}_{n=1}^{N-1}$
$\boldsymbol{t} \leftarrow \{t_n\}_{n=1}^{N-1}$
**Return** $(\tau, \boldsymbol{t})$
```

Notation

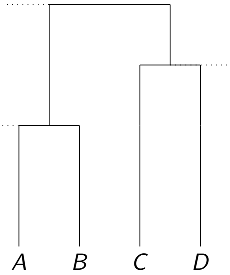
- $\{W_n, Z_n\}$ is a bipartition (a.k.a. subsplit)
  - This means the n-th coalescent event involves a clade with leaves W_n, and a clade with leaves Z_n. ($W_n, Z_n \subseteq \{x_1, \ldots x_N\}$, $W_n \bigcap Z_n = \varnothing$)

- $\{\{W_n, Z_n\}\}_{n=1}^{N-1}$ is a sequence of bipartitions
  - The topology of a tree with N leaves can be uniquely described by $N - 1$ bipartitions
  - $\#W_1 = \#Z_1 = 1$, $W_{N-1} \bigcup Z_{N-1} = \{x_1, \ldots, x_n\}$

- For $n \leq N - 1$, w and z have coalesced by the n-th event if there is a bipartition W_m, Z_m with $w \in W_m, z \in Z_m$ and $m \leq n$

Single-linkage clustering (cont)

$$T = \begin{array}{c c} & \begin{array}{cccc} A & B & C & D \end{array} \\ \begin{array}{c} A \\ B \\ C \\ D \end{array} & \begin{pmatrix} * & \mathbf{2} & 8 & \mathbf{4} \\ * & * & 4.5 & 7 \\ * & * & * & \mathbf{3} \\ * & * & * & * \end{pmatrix} \end{array}, \quad T = \begin{array}{c} \begin{array}{cccc} A & B & C & D \end{array} \\ \begin{pmatrix} * & \mathbf{2} & 5 & 6 \\ * & * & \mathbf{4} & 7 \\ * & * & * & \mathbf{3} \\ * & * & * & * \end{pmatrix} \end{array},$$



$t_3 = 4$

$W_3 = \{A, B\}$

$Z_3 = \{C, D\}$

$t_1 = 2$

$W_1 = \{A\}$

$Z_1 = \{B\}$

$t_2 = 3$

$W_2 = \{C\}$

$Z_2 = \{D\}$

Example: Two matrices T result in the same phylogenetic tree after running single-linkage clustering. Entries of T that trigger coalescence are bolded

Related work: GeoPhy

- Associate i-th taxa with a point x_i in k-dimensional space
- The variational family is a distribution q on $X \in \mathbb{R}^{N \times D}$
- Given a draw X, the corresponding tree is found by forming the distance matrix T such that $t^{(i,\, j)}$ is the Euclidean distance between x_i and x_j, and then running single-linkage clustering
- It is difficult to find a closed form for the distribution on tree topologies induced by q, so the variational family is represented by samples
- The method is generalized to points on a hyper-sphere, with distances given by geodesics

— *T. Mimori, M. Hamada. NeurIPS 2023*

Related work: VBPI

. . . Variational Bayesian Phylogenetic Inference

- Run an initializing MCMC in BEAST
- Record all subsplits (bipartitions) appearing in the initializing MCMC run
- Form subsplit Bayesian network (SBN) from these subsplits
- The variational family is a distribution on the SBNs
- Excellent accuracy in likelihood estimation
- May not be scalable, as number of subsplits in the MCMC run is large

— *C. Zhang, F.A. Matsen IV. ICLR 2019*

VIPR: Variational inference with products . . .

. . . over bipartitions

- Like GeoPhy, we consider single-linkage clustering to map from distance matrices to a tree
- We place the variational family as a distribution directly on the distance matrix (each off diagonal entry of the upper triangle is log normal)
- We derive the distribution of a tree implied by the variational family
- This distribution has a closed form as a sum over bipartitions (subsplits)

— E. Sidrow, A. Bouchard-Côté and L.T. Elliott. ICML 2025

The probability of a tree has a closed form . . .

. . . in terms of the distribution of the distance matrix T

Proposition. If the random variables $t^{\{u,v\}}$ are mutually independent, and all $q_\phi^{\{u,v\}}$ are continuous in ϕ and t for all $\{u,v\}$ with $u,v \in \mathcal{X}$, and $Q_\phi^{\{u,v\}}$ is the survival function of $t^{\{u,v\}}$, then $q_\phi(\tau, \boldsymbol{t})$ has the following form:

$$q_\phi(\tau, t) = \prod_{n=1}^{N-1} \left(\left(\sum_{\substack{w \in W_n \\ z \in Z_n}} \frac{q_\phi^{\{w,z\}}(t_n)}{Q_\phi^{\{w,z\}}(t_n)} \right) \prod_{\substack{w \in W_n \\ z \in Z_n}} Q_\phi^{\{w,z\}}(t_n) \right).$$

Here W_n, Z_n is the bipartition induced by the n-th coalescent, τ is the tree topology, and t are the coalescent times

Intuition

$$\left(\sum_{\substack{w\in W_n\\ z\in Z_n}} \frac{q_\phi^{\{w,z\}}(t_n)}{Q_\phi^{\{w,z\}}(t_n)}\right)\prod_{\substack{w\in W_n\\ z\in Z_n}} Q_\phi^{\{w,z\}}(t_n) = \sum_{\substack{w\in W_n\\ z\in Z_n}} \left(q_\phi^{\{w,z\}}(t_n) \prod_{\substack{w'\in W_n\\ z'\in Z_n\\ \{w',z'\}\neq\{w,z\}}} Q_\phi^{\{w,z\}}(t_n) \right)$$

- If two clades coalesce at time t_n, one pair (with one taxa from one clade, the other taxa from the other clade) must coalesce at time t_n in the distance matrix. And all other such pairs must coalesce after time t_n in the distance matrix (survival function $Q(\cdot) = \Pr(rv \geq \cdot)$). This marginalizes which pair coalesced
- Since entries in the distance matrix are independent, the above marginalization is summed over coalescent events

Complexity

$$\prod_{n=1}^{N-1}\left(\left(\sum_{\substack{w\in W_n\\ z\in Z_n}}\frac{q_\phi^{\{w,z\}}(t_n)}{Q_\phi^{\{w,z\}}(t_n)}\right)\prod_{\substack{w\in W_n\\ z\in Z_n}}Q_\phi^{\{w,z\}}(t_n)\right).$$

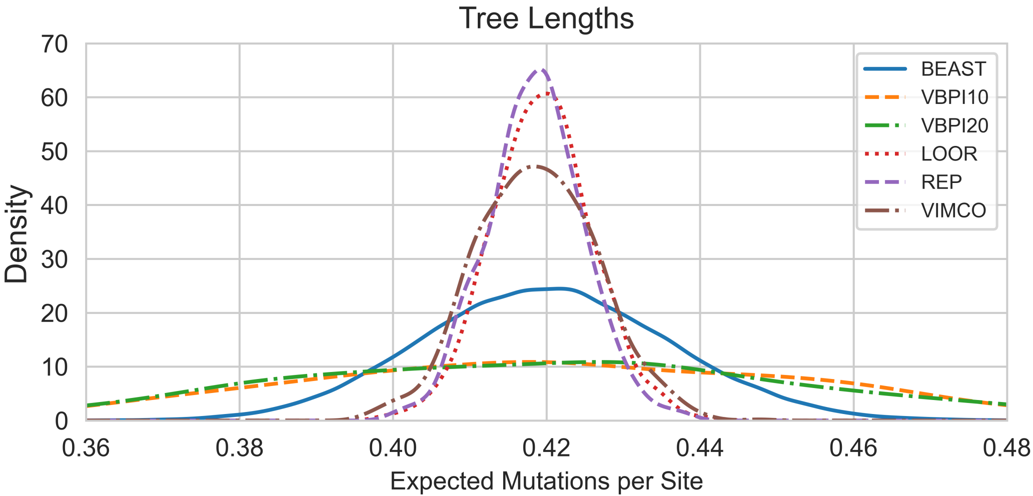
- Each pair w, z occurs once in one of the inner sums, and once in one of the inner products: $\mathcal{O}(N^2)$ operations. There are $N-1$ terms in total, so an additional $\mathcal{O}(N)$ operations
- Total complexity for evaluating q or derivatives of q: $\mathcal{O}(N^2+N)=\mathcal{O}(N)$

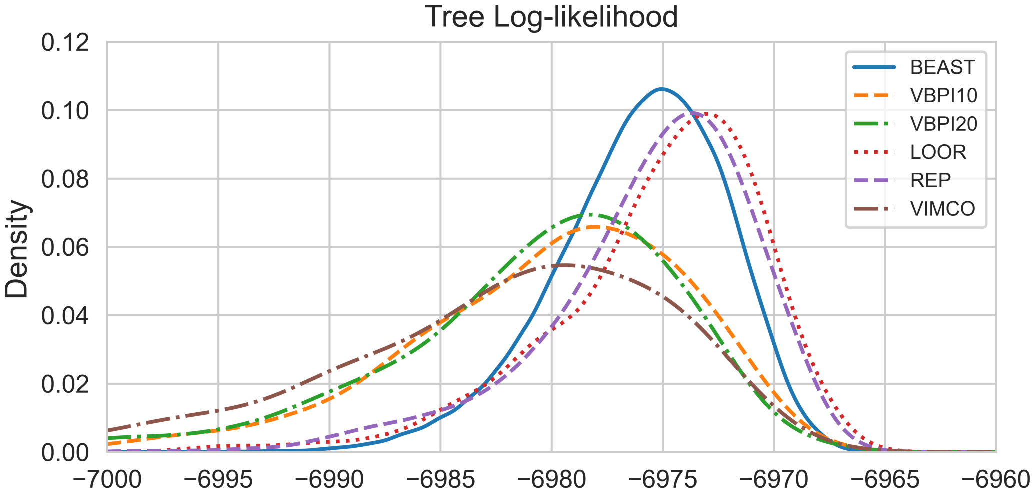
Inference

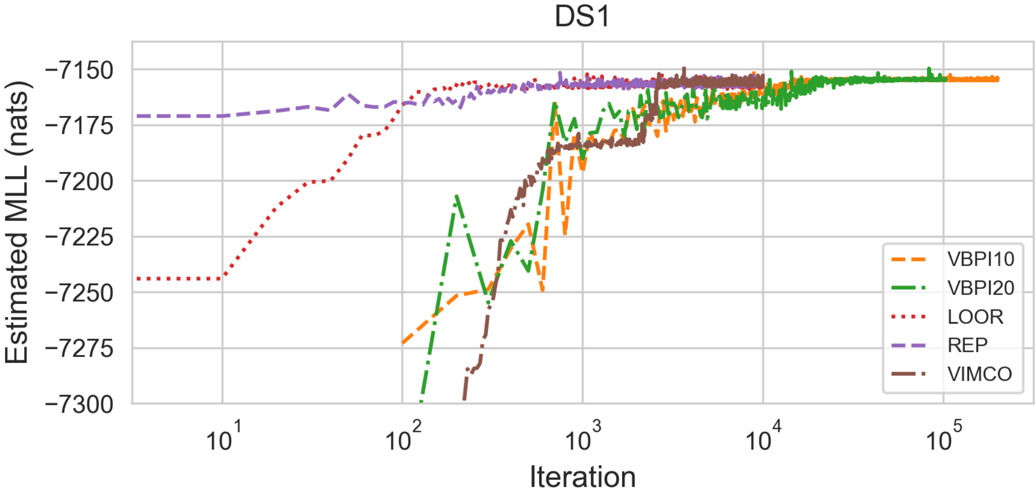
- Derivatives of the sum product in the proposition are found using automatic differentiation (torch/autodiff)

- ELBO optimization is done using:
  - VIMCO (Mnih, Rezende 2016)
  - LOOR
  - or the reparameterization-trick

Results: DS1 ($N = 27, M = 1949$)

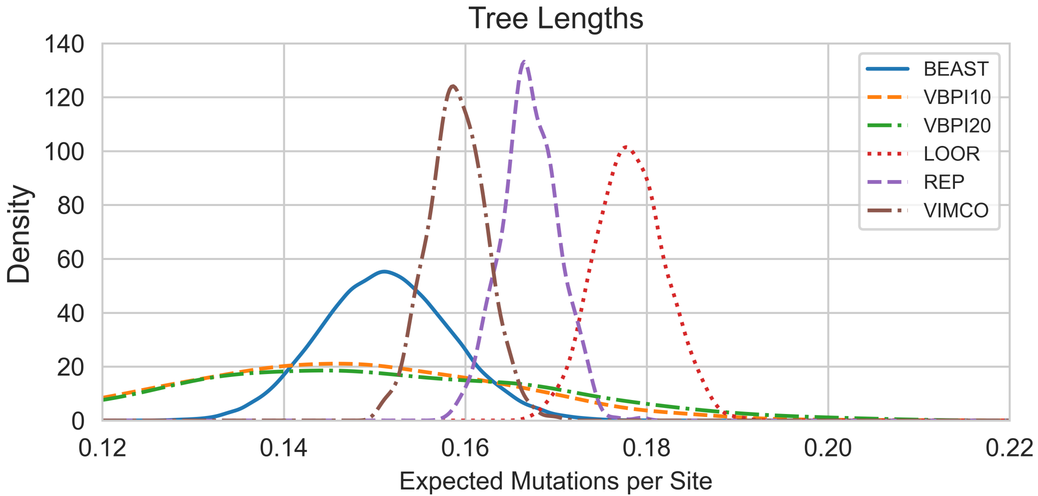


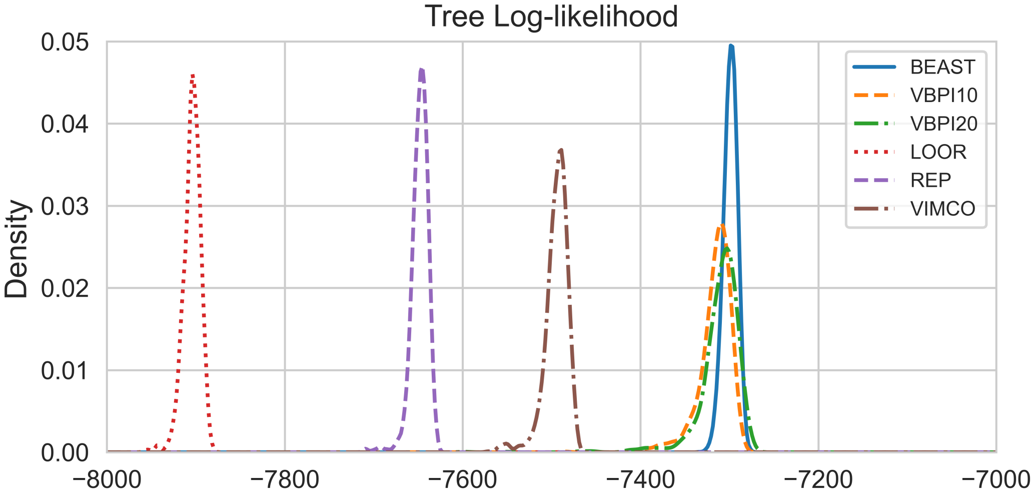

Tree Log-likelihood
Density
0.12
0.10
0.08
0.06
0.04
0.02
0.00
−7000
−6995
−6990
−6985
−6980
−6975
−6970
−6965
−6960
BEAST
VBPI10
VBPI20
LOOR
REP
VIMCO


DS1
Estimated MLL (nats)
−7150
−7175
−7200
−7225
−7250
−7275
−7300
10^1
10^2
10^3
10^4
10^5
Iteration
VBPI10
VBPI20
LOOR
REP
VIMCO

Thank You!

Results: COV ($N = 72, M = 3101$)




Tree Log-likelihood
Density
0.05
0.04
0.03
0.02
0.01
0.00
−8000
−7800
−7600
−7400
−7200
−7000
BEAST
VBPI10
VBPI20
LOOR
REP
VIMCO

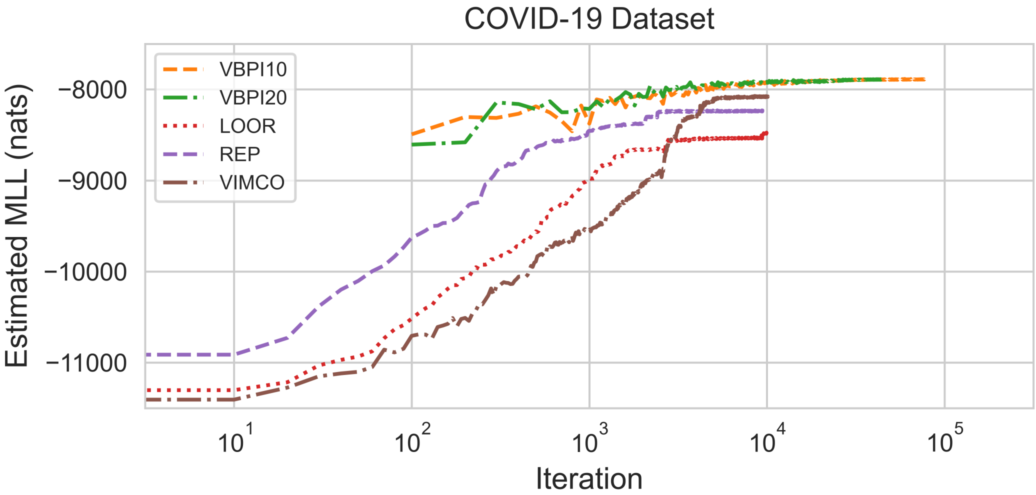

COVID-19 Dataset
Estimated MLL (nats)
−8000
−9000
−10000
−11000
Iteration
10^1
10^2
10^3
10^4
10^5
VBPI10
VBPI20
LOOR
REP
VIMCO

Table 3. ***Number of tree structure parameters versus number of taxa (Ntaxa) on simulated data with 1,000 sites.***

Ntaxa	VBPI	VIPR
8	4	56
16	44	240
32	55	992
64	3,826	4,032
128	29,939	16,256
256	127,217	65,280
512	319,533	261,632